

Multi-model seasonal forecasts for the wind energy sector

Doo Young Lee^{1,2}, Francisco J. Doblas-Reyes^{1,3}, Verónica Torralba¹ and Nube Gonzalez-
Reviriego¹

¹Barcelona Supercomputing Center (BSC), Barcelona, Spain

²Los Alamos National Laboratory (LANL), Los Alamos, USA

³Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain

Submitted to Climate Dynamics

December 2017

Revised on October 28, 2018

*Corresponding author: Doo Young Lee, Earth Sciences Department. Barcelona Supercomputing
Center (BSC), C/Jordi Girona, 29, 08034 Barcelona, Spain. E-mail: dylee1220@gmail.com

Abstract

An assessment of the forecast quality of 10m wind speed by deterministic and probabilistic verification measures has been carried out using the original raw and two statistical bias-adjusted forecasts in global coupled seasonal climate prediction systems (ECMWF-S4, METFR-S3, METFR-S4 and METFR-S5) for boreal winter (December-February) season over a 22-year period 1991–2012. We follow the standard leave-one-out cross-validation method throughout the work while evaluating the hindcast skills. To minimize the systematic error and obtain more reliable and accurate predictions, the simple bias correction (SBC) which adjusts the systematic errors of model and calibration (Cal), known as the variance inflation technique, methods as the statistical post-processing techniques have been applied. We have also built a multi-model ensemble (MME) forecast assigning equal weights to datasets of each prediction system to further enhance the predictability of the seasonal forecasts. Two MME have been created, the MME4 with all the four prediction systems and MME2 with two better performing systems. Generally, the ECMWF-S4 shows better performance than other individual prediction systems and the MME predictions indicate consistently higher temporal correlation coefficient (TCC) and fair ranked probability skill score (FRPSS) than the individual models. The spatial distribution of significant skill in MME2 prediction is almost similar to that in MME4 prediction. In the aspect of reliability, it is found that the Cal method has more effective improvement than the SBC method. The MME4_Cal predictions are placed in close proximity to the perfect reliability line for both above and below normal categorical events over globe, as compared to the MME2_Cal predictions, due to the increase in ensemble size. To further compare the forecast performance for seasonal variation of wind speed, we have evaluated the skill of the only raw MME2 predictions for all seasons. As a result, we also find that winter season shows better performance than other seasons.

47 **Keywords**

48 seasonal prediction systems; statistical post-processing; multi-model ensemble; 10m wind

49 speed; forecast verification

50

51

1. Introduction

Modern society is looking forward to the growth and widespread diffusion of renewable energies such as wind and solar power, hopefully contributing to the major part of the world energy supply (Frankfurt School-UNEP Collaborating Centre 2016). Wind power will especially play an increasingly important role in providing a substantial share of renewable energy supply over the coming years (Troccoli et al. 2010). The ability to anticipate and respond to changes in wind energy supply and demand is essential to stabilize and secure the entire electricity network. For this reason, accurate and reliable information from weather and climate forecasts is required, for the development and use of wind energy (Troccoli et al. 2010, Vladislavleva et al. 2013).

Previous works have dealt with the sensitivity of the energy system to the variability at either short or long time scales, such as weather forecasts (Amin 2013, Foley et al. 2012, Troccoli et al. 2013, Vladislavleva et al. 2013) or climate change projections (Ebinger and Vergara 2011, IPCC 2012, Koletsis et al. 2016), while there are only a few very recent studies on the use of seasonal climate forecasts for wind energy applications (Clark et al. 2017 and Torralba et al. 2017).

In the last few years, even though the performance of the seasonal climate prediction has been significantly improved, systematic errors still remain (Feddersen et al. 1999, Wang et al. 2008, Kug et al. 2008, Alessandri et al. 2010). Many climate scientists and climate services communities have tried to deal with the problems, such as model error and forecast uncertainty, for producing better seasonal climate forecast information relevant to user applications (Buontempo et al. 2014, Coelho and Costa 2010, Morse et al. 2005, Palmer et al. 2005).

The main aim of the present study is the assessment and improvement of forecast quality and accuracy of seasonal climate prediction system in predicting global wind speed. To this

end, several deterministic and probabilistic verification measures were applied to evaluate the quality of individual forecast systems and multi-model ensemble (MME) against reanalysis dataset.

To reduce the forecast uncertainty and improve reliability of forecasts for the seasonal wind speed, a statistical post-processing stage using two bias-adjustment techniques, simple bias correction (Pan & den Dool 1998, Leung et al. 1999, Acharya et al. 2013) and calibration (Doblas-Reyes et al. 2005, Johnson and Bowler 2009, Charles et al. 2011, Torralba et al. 2017), has been applied.

Furthermore, systematic assessment of the MME, based on the combination with equal weights of four different independent forecast systems, has been also carried out for the purpose of the enhancement of seasonal predictability for wind energy sector and satisfying the needs of the wind-energy community.

The rest of this paper is organized as follows. Section 2 presents a brief description of the seasonal prediction systems and reanalysis dataset used as a reference, statistical post-processing methods, and verification measures used in this study. The results from the forecast quality assessment of the prediction systems and the MME are described in Section 3. The summary and main conclusions are given in the final section.

2. Data and methodology

In this study we have focused on the quality assessment of the seasonal mean 10m wind speed, as one of the key variables to wind power supply, in the common period (1991 to 2012) between all data sets of four coupled global seasonal prediction systems. See Section 2.1 and Table 1 for the prediction systems.

To derive a more accurate estimate of model prediction performance and avoid overfitting, the forecast and observed anomalies are obtained from the standard “leave-one-out” cross-

validation method (Michaelsen 1987, Jolliffe and Stephenson 2003). This cross-validation method computes seasonal mean anomalies for each model, from the corresponding seasonal climatology obtained by excluding seasonal mean data from the target year.

A land-sea mask is also applied to represent information over land only. Sea points with a depth equal to or less than 50m are included on land to consider offshore wind farms installed in the relatively shallow ocean worldwide.

2.1 Forecast systems

Seasonal predictions of 1-month lead-time initialized on 1st November (December through February, DJF) performed by four coupled global seasonal prediction systems: the European Centre for Medium-Range Weather Forecasts seasonal forecast system 4 (ECMWF-S4; Molteni et al. 2011), Météo-France's System 3 (METFR-S3; Madec et al. 1998, Déqué et al. 1999, Royer et al. 2002, Daget et al. 2009, Weisheimer et al. 2009, Chevallier and Salas-Mélia 2012), Météo-France's System 4 (METFR-S4; Voldoire et al. 2013) and Météo-France's System 5 (METFR-S5; Meteo France 2015a, b) have been analysed over the 1991-2012 period. These prediction systems have been selected by taking into consideration the availability of 6-hourly seasonal forecasts for 10m wind speed over the period 1991-2012.

The ECMWF-S4 consists of the ECMWF Integrated Forecast Model (IFS) and Nucleus for European Modelling of the Ocean (NEMO) as atmospheric and oceanic components, respectively. Its hindcast (historical forecast) has 51 ensemble members (simulations) and the standard forecast time horizon is seven-month, initialized on the 1st day of every month from 1981 until 2010. Details for the ECMWF-S4 can be found in Molteni et al. (2011). The METFR-S3 utilizes the fourth version of the Action de Recherche Petite Echelle Grande Echelle (ARPEGE; Déqué et al. 1999, Royer et al. 2002) as atmospheric component. The ocean component is the global version of the Océan PARallélisé (OPA; Madec et al. 1998,

Daget et al. 2009) model version 8.2. Its hindcast has 11 ensemble members, all starting on the 1st day of every month. Simulations are seven-months long and cover the period 1981-2012. The METFR-S4 has been running operationally since September 2012. It consists in a 15 ensemble members hindcast starting once per month over 1991-2012 based on ARPEGE-Climat version 5.2 coupled with NEMO 3.2. In early 2015, the METFR-S5 was introduced with an ARPEGE version 6.1(T255 L91) as atmospheric model and the NEMO version 3.2 with a 1-degree horizontal resolution and 42 vertical levels as oceanic model. It accounts for 15 members and spans 22 years from 1991 until 2014. See Table 1 for a brief description of the systems.

2.2 Observed dataset

For the forecast verification, we have used the ERA-Interim reanalysis (Dee et al. 2011). ERA-Interim is ECMWF's most recent atmospheric reanalysis, covering the modern satellite era from January 1979 to the present. It is based on a 2006 version of the ECMWF IFS and utilizes a four-dimensional variational analysis (4D-Var) for data assimilation. The spatial resolution of the data set is approximately 80 km (T255 spectral) on 60 vertical levels from the surface up to 0.1 hPa.

2.3 Methodology

2.3.1 Post-processing methods

To improve aspects of the forecast quality by reducing the impact of the model systematic errors, two post-processing methods are employed. The simple bias correction (SBC) method is known as a standardized reconstruction technique which adjusts the systematic errors of the model using the standardized anomaly of the ensemble mean. By default, standardized anomalies of the ensemble mean are measured by subtracting the climatology of the

ensemble mean and normalizing with the standard deviation of ensemble mean. To estimate the bias adjusted forecast, the standardized anomaly of the ensemble mean is reconstructed by multiplying the observed standard deviation and adding the observed climatology (Pan and van den Dool 1998, Leung et al. 1999, Acharya et al. 2013, Torralba et al. 2017).

In the calibration (Cal) method, we use the variance inflation technique that has been proposed in several studies (Doblas-Reyes et al. 2005, Johnson and Bowler 2009, Charles et al. 2011, Torralba et al. 2017). It assumes that the bias adjusted ensemble forecasts by Cal method should have the same climatological variance as observations. To obtain the inflated ensemble member as more reliable ensemble prediction, the inflation of both the ensemble mean and the ensemble spread (as the difference of ensemble member with the ensemble mean) is required. Coefficients, including the variance inflation, of the ensemble mean and spread are computed with observed standard deviation, ensemble mean standard deviation, correlation between observation and ensemble mean, and square root of the mean variance of the ensemble spread. For more detailed information on the method, the readers are referred to the above-mentioned papers related to calibration.

In this study, to analyze each different behavior of bias adjusted forecasts in skill quality assessment, we have applied these two post-processing methods (SBC and Cal) to the seasonal hindcasts of individual models and MME.

2.3.2 Multi-model ensemble

Many studies have reported that the multi-model ensemble (MME) among the results of various prediction models, considering the performance of each model, can produce much more accurate and reliable forecasts (Kharin and Zwiers 2002, Peng et al. 2002, Hagedorn et al. 2005, Min et al. 2009, Weigel et al. 2010). The MME techniques are known as a useful and practical approach for reducing the inherent errors contained in individual models and

providing better performance than the constituent individual models (Krishnamurti et al. 2000, Palmer 2000, Pavan and Doblas-Reyes 2000, Peng et al. 2002, Hagedorn et al. 2005, Doblas-Reyes et al. 2005, Yun et al. 2005, Weigel et al. 2008, Min et al. 2009, Lee et al. 2011, 2013, 2015, Jeong et al. 2012, 2015).

As a deterministic MME approach, we use simple arithmetic mean for combining multi-model seasonal predictions based on the different prediction systems. In this technique, equal weights are assigned to the ensemble mean predictions of each of the prediction systems.

$$E = \bar{O} + \frac{1}{N} \sum_{i=1}^N (F_i - \bar{F}_i) \quad (1)$$

where, E is the multi-model ensemble mean prediction from the different models. \bar{O} is the observed climatology from ERA-Interim over the training period. N is the number of prediction systems. F_i is the i^{th} ensemble mean forecast out of N prediction systems. \bar{F}_i is the seasonal climatology of ensemble mean forecast. The MME results have been computed in cross-validation for the raw, simple bias corrected and calibrated data.

For the probabilistic MME analyses, forecast probabilities for each tercile (above normal (AN), near normal (NN), and below normal (BN)) category are estimated separately for each individual prediction system, and then such probabilities for each category are combined by applying the simple average with equal weights.

$$P(E_j) = \frac{1}{N} \sum_{i=1}^N P(E_j | M_i) \quad (2)$$

where, P is a forecast probability for each j -event, E_j is j -event (i.e., either AN, NN or BN), M_i is i -model, and N is the number of models. In this equation, $P(E_j | M_i)$ is a forecast probability of the event conditioned on the i -model (i.e., the i -model forecast of j -event).

In this work, tercile events have been used for illustrative purposes, but other categories that are tailored to the specific needs of wind energy customers could be defined.

2.3.3 Forecast quality measures

To investigate the forecast ability of the seasonal prediction systems to reproduce adequately the observed 10m wind speed variability, a set of deterministic and probabilistic verification measures, such as the temporal correlation coefficient (TCC), fair ranked probability skill score (FRPSS) and reliability diagram, for each individual model and MME prediction are estimated over the retrospective forecast period (Jolliffe and Stephenson 2003, Wilks 2006).

The TCC is designed to analyze the spatial distribution of forecast skills between forecasts and their corresponding observations. Using a two-tailed Student's t-test, the statistical significance of the TCC at the 90% confidence level is calculated.

One of the more commonly used probabilistic measures to evaluate forecasts of multiple categories is the ranked probability skill score (RPSS; Epstein 1969, Murphy 1971, Daan 1985). The RPSS measures the cumulative squared error between the categorical forecast probabilities and observed categorical probabilities relative to a reference (Wilks 2006). When the value of RPSS equals to 1, it implies that the observed category is always predicted with 100% confidence. $RPSS = 0$ implies that the prediction skill is same as reference prediction (observed climatology, in our case) and a score < 0 means that the forecast system performs worse than climatology.

The RPSS can make unfair evaluations for inter-comparing ensemble predictions, due to the different number of ensemble members. In this regard, Ferro et al. (2008, 2014) mentioned that the RPS can be adjusted to provide a fair way in evaluating ensemble forecasts. For a fair evaluation, we have applied the fair RPSS (FRPSS) to the seasonal forecasts of not only the individual prediction systems, but also the MME. In this way, it is possible to compare forecasts with a different ensemble size. In this study, the FRPSS is

calculated for tercile events. The statistical significance of the FRPSS is computed based on the 95% confidence level from a one-tailed Z-test.

The reliability diagram shows how well the forecast probabilities correspond to the observed relative frequencies of occurrence of an event for each of the forecast tercile (AN, NN, or BN) categories (Jolliffe and Stephenson 2003, Wilks 2006). The diagonal line on the reliability diagram indicates perfect agreement between the forecast probabilities and the observed relative frequency. The horizontal line (referred to as the no-resolution line) represents the observed climatological frequency of the event, while the vertical line (referred to as the no-sharpness line) is for the climatological forecast probability. The no-skill line is defined as a line halfway between the no-resolution and perfect reliability lines. The reliability diagram is usually accompanied by a sharpness diagram as an indication of the sample size (frequency of forecasts) in each probability bin, such as a histogram. The sharpness diagram shows the tendency of the forecast to predict extreme values, i.e., a forecast of climatology means no sharpness. Vertical bars on the diagonal depict the 95% consistency bars, constructed by bootstrapping 500 samples with replacement from the original sample, for each bin of the reliability diagram (Brocker and Smith. 2007). The consistency bars allow an immediate visual interpretation of the quality of the probabilistic forecast system and also increase more credibility of interpretations of reliability diagrams.

3. Seasonal hindcast quality of the wind speed

3.1 Boreal winter

Figure 1 displays the spatial distributions of the TCC of cross-validated raw data sets from the individual models and MMEs for the 10m wind speed predictions over the global region during winter (December through February, DJF) for period 1991-2012. In general, the prediction skill of the ECMWF-S4 (Figure 1c) is significantly superior to those of the

other three individual models, even though it shows poor performance in some regions, such as the northern part of Africa, southern Europe and eastern Russia. Especially, the ECMWF-S4 seasonal forecast shows statistically significant positive coefficients over the North America, northern South America, most of maritime continent, eastern Africa and northern portion of China. The METFR-S3 (Figure 1d) shows similar features to the METFR-S4 (Figure 1e), except the significantly positive TCC in central Europe, eastern Africa and northern China. Figure 1f clearly shows that the METFR-S5 has slightly better performance than the previous versions of METFR forecast systems. We can also find that common areas showing high prediction skill from each prediction system are confined to the Maritime Continent, southern North America and northern South America. For enhancement of seasonal forecast quality and providing better performance than the constituent individual models, we have carried out the ensemble mean predictions by employing all prediction systems (MME4). Based on the research results reported by Lee et al. (2011 and 2013) showing that the skills of MME comprising of only the more skillful models are relatively better than those of a comprehensive MME which contains all the available models, we have also added the MME prediction (hereinafter MME2) by using the two best performing systems (ECMWF-S4 and METFR-S5) as current operational models. The TCC skill of MME4 prediction is considerably improved as compared to those of the individual models. However, certain limitations remain in improving the predicted 10m wind speed in some regions, particularly over some parts of South America, Africa and Australia. The significant spatial distributions of the TCC of the MME4 predictions are very close to those of the ECMWF-S4 prediction. The MME2 prediction that is generated by the best two models shows almost similar performance to the MME4 prediction. Over the southern and northeastern parts of Australia, central part of Russia and eastern portion of Africa, it can be seen that the MME2 prediction has wider distributions for significantly positive correlation

coefficients as compared to the MME4 prediction, while in the northeastern Russia, eastern Europe and southern part of Alaska, the distributions with positively significant correlation coefficients in the MME4 are a little bit wider than in the MME2.

Figure 2 shows the skill scores for probabilistic forecasts of seasonal prediction systems in terms of the cross-validated raw 10m wind speed for winter. The ECMWF-S4 generally shows a better performance in FRPSS than other prediction systems (Figure 2c). The ECMWF-S4 wind speed seasonal predictions have significantly positive skill over the United States, northern South America, northern China, and some parts of Maritime Continent. It is difficult to find distinctly negative skill areas in the ECMWF-S4. The METFR-S3 has few significantly positive regions, but in some regions of central Europe, United States and Maritime Continent. The negative skill scores for the METFR-S4 are found over large areas. In the METFR-S5, the number of significant regions with positive skill increases compared to the two previous forecast systems, METFR-S3 and METFR-S4. Especially, over the Canada and central Europe that are regions particularly relevant for the wind industry, it is found that the METFR-S5 shows good performance compared to other prediction systems. The MME4 prediction, consisting of all four models, shows a large spatial distribution with significant higher skills over the North America, northern Europe and China, and outperforms the all forecasts of the individual prediction systems. The MME2 prediction also shows that the overall performance of FRPSS is as good as the MME4 prediction, though the slightly lower skills are shown in some regions, such as Eastern Europe and central Russia.

Until now, all MME results shown have been based on using the cross-validated raw predictions from the individual models. In order to further enhance the MME forecast performance from minimizing the model systematic uncertainties and errors, we have applied two different bias-adjustment methods (SBC and Cal, refer to section 2.3.1) to the seasonal

predictions of each system. The MME predictions for each bias-correction approach are constructed to compare the two different behaviors of bias-adjustment in skill assessment.

The results of the MME based on the bias-adjusted seasonal prediction show significantly positive skills over the Indonesia, middle Europe, northern China, eastern Africa, northern South America, and most of North America (Figure 3). In Figure 3a and 3c, the TCCs for DJF 10m wind speed of simple bias corrected MME4 (MME4_SBC) and calibrated MME4 (MME4_Cal), obtained from the combination of the post-processed all individual model predictions, are indicated. The significant spatial patterns of MME4_SBC prediction are almost similar to those of the raw MME4 of Figure 1a. The MME4_Cal also shows a similar distribution to the raw MME4, but it has slightly lower skills compared to the raw MME4. In addition, the MME4_Cal has the characteristic that the spatial distribution of TCC shows more noisy patterns of a point-like shape compared with the two other types of TCCs for the raw MME4 and MME4_SBC. This might be caused by the uncertainties of the coefficients estimated from the computational process of calibration method (Torrallba et al. 2017). These same features are also found in the TCC distribution of MME2_Cal using the two best performing models. In Figure 3b, it is shown that the MME2_SBC has nearly the same pattern and performance compared to the MME4_SBC, except for a little difference in the Europe, Russia and Canada. The significantly positive skill distributions of the MME2_Cal of Figure 3d are, in general, similar to those of the MME4_Cal. One interesting feature of the skill distribution is that both the bias-adjusted MME2_SBC and Cal show improved performance in the northern China and eastern part of Africa as compared to the corresponding two MME4s.

We have also calculated the root mean square skill score (RMSSS, Murphy 1988), a basic non-dimensional measure of the strength of the linear relationship between forecasts and observations based on root mean square error values, of all MME predictions for the raw and

two bias-corrected datasets to check the deterministic forecast accuracy with respect to the observed climatology (Figure S1). The relatively high levels of skill of the all MME predictions are commonly distributed over the North America, northern South America, and Indonesia region. Particularly, the significantly positive RMSSS by a one-tailed F-test tends to only appear in the Indonesia region. The distinctly positive skill distributions (of more than 0.1) of all MME2 predictions are much wider than the corresponding distribution of all MME4 predictions.

To further compare the probabilistic forecast accuracy of the bias-adjusted MMEs, the FRPSS for 10m wind speed has been computed (Figure 4). The significant positive values in FRPSS of MME4_SBC prediction are found over the North America, northern South America, northern Europe, central Russia, eastern Africa, and northern China. The MME4_SBC prediction has almost similar spatial distributions to the raw MME4 prediction of Figure 2a. In the MME4_Cal of Figure 4c, the regions with significant distributions are almost same as those in MME4_SBC prediction, while the spread of the regions is much less extensive than that in MME4_SBC prediction. As compared to the MME4_SBC prediction, the MME2_SBC, almost similar to MME2 raw prediction, has no substantial change in the significant skill, except for the distinct differences at the Eastern Europe and central Russia. The significant skill patterns of both the MME2_Cal and MME4_Cal predictions look quite similar to each other. Over the northern Europe and northwestern China regions, the FRPSSs of the MME2 predictions taken by the both bias-adjustment methods show noticeable differences. The significant FRPSS areas of MMEs by using the SBC method in Figure 4 are relatively more widely distributed as compared to the corresponding areas of MMEs by using the Cal method.

Figures 5 displays the reliability diagrams (described in section 2.3.3) for the two bias-adjusted MME4 (MME4_SBC and Cal) and MME2 (MME2_SBC and Cal) categorical

probability forecasts of the above (top) and below (bottom) normal 10m wind speed in the globe, respectively. In Figure 5a and 5c, the reliability curves in the MME4_Cal predictions are much closer to the diagonal than those in the MME4_SBC predictions and indicate an almost perfect reliability shape for the both categorical events. Several studies (Doblas-Reyes et al. 2005, Charles et al. 2011, Torralba et al. 2017) also found that calibrated probabilistic forecasts show significant improvements in the reliability of the forecasts. On the other hand, the curves for the of MME2_Cal predictions show the similar reliability patterns to those for the MME2_SBC predictions (Figure 5b and 5d). As compared to the MME4 predictions of Figure 5a and 5c, the MME2 predictions of Figure 5b and 5d have the less reliable shapes in both the above and below normal events. Especially in the comparison of Cal method rather than SBC method, the difference of reliability is much more clearly shown. This issue between MME4 and MME2 prediction may be caused by the different size of total ensemble members which are combined to build the both MME probability forecasts. Though the two individual prediction systems, such as METFR-S3 and METFR-S4, not employed in the MME2 predictions have a considerably poor performance in reliability diagram (see Figure S2), it is shown that an increased total ensemble size plays a very important role in the estimation of reliability (Richardson 2001, Hagedorn et al. 2005). The sharpness diagrams, the number of probability forecasts falling into each probability bins, at the right of reliability diagram in Figure 5 are plotted. The frequencies of MME2_Cal and MME4_Cal forecast probabilities are larger than those of MME2_SBC and MME4_SBC predictions in those bins centered close to the climatological probability. This means that the MME_Cal predictions have a smaller sharpness than the MME_SBC predictions.

To further understand the effect of the multi-model approach on forecast performance as measured by the reliability for the probability forecast, we have investigated the reliability for the probabilistic categorical forecasts in terms of the raw predicted dataset of individual

models and their MME (Figure S2). It is found that the reliability shapes of MME4 raw predictions are almost similar to those of MME4_SBC predictions in Figure 5 for above and below normal categories. We have already mentioned this fact in the estimation of the forecast performance verified by the deterministic and probabilistic measures, such as TCC, RMSSS and FRPSS. The probabilistic forecast of ECMWF-S4 depicts the more reliable pattern than other three individual prediction systems. The reliability curve of the METFR-S3 prediction tends to be close to the climatological observed frequency line in the both categories. Richardson (2001), who examines the effect of ensemble size on the reliability diagram, mentioned that the reliability of forecast probability for ensemble prediction system can strongly depend on the number of ensemble members used. However, even though the two prediction systems (METFR-S4 and S5) hold the ensemble members of the same size, the METFR-S5 prediction system shows a better reliability than the METFR-S4 prediction system. As reported by many researchers (Hagedorn et al. 2005, Langford and Hendon 2013, Kirtman et al., 2014), it is distinctly shown that the MME prediction outperforms individual models' predictions in the aspects of reliability of a probabilistic forecast. The sample size in each forecast probability bin, as a histogram, is also indicated in Figure S2. It can be discerned that the frequencies of forecasts for MME4 in extreme bins are lower than those for individual models, while in the climatological probability bins the frequencies of MME4 forecast are larger than those of probability forecasts for each system, as noted by many studies (Kharin et al., 2009, Yang et al., 2016, Barnston et al., 2003, Kirtman et al., 2014). This indicates that the MME probability forecast has lower sharpness for forecasts of extreme values than the probability forecasts for the individual models.

Reliability diagrams for the Northern Europe (NEU: 15°W-45°E, 45°N-75°N), where it is one of the regions showing the significant FRPSS in the raw MME predictions (see Figure 2) and there are many areas of interest for wind energy, are shown in Figure 6. In the below

normal category event, the MME4 and MME2 predictions tend to have the slightly steeper slopes than the diagonal line in the right-hand side beyond the climatological frequency, while in the above normal event, they show the gentle slopes. The reliability curves of the MME predictions for both the bias-adjustment methods in the NEU region show the narrow ranges, with the values from 0 to 0.7 for MME4 and 0 to 0.8 for MME2, for forecast probability compared to the corresponding curves in the global region. There is little difference in the reliability lines between the MME4_SBC and MME4_Cal predictions for both categorical events, except for a difference in the last bin of curves for the above normal category. For the MME2 predictions, the curves of the SBC adjustment method show the similar reliable patterns to those of the Cal adjustment method in the above and below normal categories, respectively. As compared to Figure 5, this result shows that the reliability diagnosis is also greatly influenced by the number of total forecasts, such as grid points, obtained from selected region as well as a given total ensemble size from each model. As for the sharpness diagrams, similarly to those in global region, the numbers of MME4_Cal probability forecast are much larger than those of MME4_SBC forecasts in the bins centered close to the climatological probability, but on the contrary to globe, in terms of the MME2 predictions in the NEU, the numbers of the probability forecast by the SBC method are much larger than by the Cal method.

3. 2 Other seasons

In the previous section, we have focused on the comparison of performance between the MME4 using all predictions and MME2 using the two best performing predictions, as well as the individual model predictions in terms of the raw and bias corrected datasets, for only winter (DJF) season. Hence, in this section, the spatial distributions of FRPSS for the only raw MME2 prediction in the four seasons (DJF, March to May; MAM, June to August; JJA,

and September to November; SON) are analyzed (Figure 7). The MME2 prediction for winter (DJF) has the significant spatial distribution over the North America, northern South America, northern Europe, China and eastern Africa. The MME2 prediction of spring (MAM) wind speed shows the spatial patterns of the significant positive FRPSS in the central United States, central parts of South America and Africa, southern China, and western portion of Australia. In summer (JJA), the significant spatial patterns of the FRPSS for the MME2 prediction are mostly concentrated over the tropical region of 20°S-20°N, especially over the Maritime Continent and Indian subcontinent. The MME2 prediction in autumn (SON) has no significant spatial distributions in the U.S., northern China and eastern Russia, and most of significant patterns tend to appear in the tropics (20°S-20°N) and Southern Hemisphere (20°S-90°S). Generally, the MME2 prediction over the Northern Hemisphere (20°N-90°N) in DJF season shows a better performance than the corresponding predictions in other seasons.

In Figure S3, we have also examined the TCC based on the raw MME2 prediction in all four seasons to further compare the performance of deterministic prediction for wind speed variation. The spatial distribution patterns of the significantly positive TCC are almost similar to those of the significant FRPSS for all seasons. In DJF and MAM seasons, the significant skills generally appear over the North America and northern China. The significantly positive TCCs in the central Europe are only found in winter season. In the northern South America, eastern Africa and Maritime Continent, the significant positive skills are always distributed for all four seasons. Similarly to skill distributions in FRPSS, winter season generally shows an even higher performance over the Northern Hemisphere compared to other seasons.

4. Summary and conclusions

The forecast ability of global coupled seasonal climate prediction systems (ECMWF-S4, METFR-S3, METFR-S4 and METFR-S5), selected by the availability of 6-hourly seasonal

forecasts for 10m wind speed, has been investigated to provide more useful and reliable climate information that can be used for the wind energy industry. We have first carried out the assessment of the wind speed forecast quality by the deterministic and probabilistic verification measures for winter (DJF) season over the 22 years period 1991–2012 using the corresponding wind speeds from the ERA-Interim reanalysis. To avoid overfitting of retrospective forecasts, we used the leave-one-out cross-validation for each target year of the study period, and then two statistical post-processing techniques, such as SBC and Cal, have been applied to the original raw forecasts to reduce the systematic model bias and improve the reliability and accuracy of forecasts. Using the MME approach assigning equal weights to datasets of each forecast system, we have also tried to further enhance the predictability of the seasonal forecasts. In this study, the two combinations of seasonal MME predictions named as the MME4 (employing all seasonal prediction systems) and MME2 (employing two better performing seasonal prediction systems) have been carried out.

For DJF 10m wind speed, the ECMWF-S4 prediction system generally showed the better performance in the global geographical distributions of the TCC and FRPSS than other prediction systems, except for northwestern Canada, central Europe and some parts of Australia. The latest version of METFR forecast system showed considerably improved performance compared to the previous versions. The MME4 prediction indicated consistently higher TCC and FRPSS than the individual models, even though there still remains room for skill improvement in some regions. The significant skill regions of MME2 prediction are almost similar to MME4 prediction, which is feature that has also been found in skill assessment of the bias-adjusted MME predictions. The MME predictions based on the simple bias correction (SBC) method showed considerably similar skill patterns to those by calibration (Cal) method, but the significant MME skill areas obtained from SBC method were more spread out as compared with those from Cal method.

The bias adjusted MME4 prediction based on the calibration method (MME4_Cal), unlike MME4_SBC, showed an almost perfect reliability for above and below normal categorical events over globe. However, in the MME2_Cal prediction obtained from removing the two prediction systems (e.g., METFR-S3 and METFR-S4) that have shown the poor performance, it was difficult to get the effective improvement on reliability compared to the MME2_SBC prediction. This fact shows that an increase in ensemble size, though the two less skillful systems abovementioned are employed, would work much more effectively on improvement of the reliability that is especially based on calibration method. In addition, comparison of the reliability between global and the local areas (e.g., not only NEU from Figure 6 but also other regions such as North America (Figure not shown)) in terms of the bias-adjusted MME4 and MME2 predictions implies that the size of the selected area would be one of the factors that may influence reliability diagram.

Based on the forecast performance of the MME2 predictions showing quite similar performance to the MME4 predictions in the aspects of forecast quality, we have further examined seasonality of the MME2 raw prediction using the FRPSS and TCC as forecast verification measures. As a result, it has been revealed that the MME2 raw predictions in 10m wind speed generally have high skills in aspects of probabilistic and deterministic predictions over the Northeastern China during DJF, Maritime Continent and India subcontinent for JJA, central China and West Asia for MAM, and southern Australia for SON season.

This study proves that the MME approach is very practical for providing useful seasonal climate information to wind energy community and furthermore, the skill enhancement of individual prediction systems with the adequate ensemble size is crucial to improve the MME seasonal prediction. In addition, the statistical bias-adjustment method, especially calibration method, plays an important role in providing information of the improved reliability.

498 The present study, however, is subject to the limitations of the number of prediction
499 systems that have seasonal forecasts of 6-hourly 10m wind speed available. Nonetheless, this
500 forecast quality assessment demonstrates the possibility of providing better climate
501 information for global wind speed to improve the current sources of information used in wind
502 energy applications and decision-making at the seasonal time scale.

503 Finally, this study has been carried out with focus only on seasonal wind speed as part of
504 the project base. However, since wind energy production is closely related to wind direction
505 as well as speed, further investigation combining the predictability of wind direction would
506 be necessary to provide more useful information for the wind energy sector. Furthermore,
507 more detailed analyses on the forecast performance of seasonal wind speed with different
508 forecast lead times also need to be done to further elucidate how the forecast quality of long-
509 lead seasonal prediction can greatly impacts the long-term planning of wind power generation.

510

Acknowledgments

This research was funded by the Spanish Ministry of Economy (MINECO) under the framework of the RESILIENCE project (CGL2013-41055-R). We are grateful to ECMWF and Météo France as the supporting institutions that provide with climate prediction datasets.

References

- Acharya, N. et al. (2013) On the bias correction of general circulation model output for Indian summer monsoon. *Meteorol Appl* 20:349–356
- Acharya, N. et al. (2014) Prediction of Indian summer monsoon rainfall: A weighted multi-model ensemble to enhance probabilistic forecast skills. *Meteorol Appl* 21:724–732
- Alessandri, A. et al. (2010) The INGV–CMCC seasonal prediction system: Improved ocean initial conditions. *Mon Weather Rev* 138:2930–2952
- Amin, M. (2013) Energy: The smart-grid solution. *Nature* 499:145–147
- Barnston, A. G., S. J. Mason, L. Goddard, D. G. DeWitt, and S. E. Zebiak (2003) Multimodel ensembling in seasonal climate forecasting at IRI. *Bull Am Meteorol Soc* 84:1783–1796. doi:10.1175/BAMS-84-12-1783
- Brocker, J., and L.A. Smith (2007) Increasing the reliability of reliability diagrams. *Weather Forecast* 22:651–661
- Buontempo, C. et al. (2014) Climate service development, delivery and use in Europe at monthly to inter-annual timescales. *Climate Risk Management* 6:1–5
- Charles, A. et al. (2011) Comparison of techniques for the calibration of coupled model forecasts of Murray Darling Basin seasonal mean rainfall. CAWCR Tech Rep No. 040. http://www.cawcr.gov.au/technical-reports/CTR_040.pdf

535 Chevallier, M., and D. Salas-Mélia (2012), The role of sea ice thickness distribution in the
 536 Arctic sea ice potential predictability: A diagnostic approach with a coupled GCM, J.
 537 Clim., 25(8), 3025–3038.

538 Clark, R. T., Bett, P. E., Thornton, H. E. and Scaife, A. A. (2017) Skilful seasonal
 539 predictions for the European energy industry. *Environ. Res. Lett.* **12**, 024002.

540 Coelho, C.A.S. and S.M.S. Costa (2010) Challenges for integrating seasonal climate
 541 forecasts in user applications. *Curr Opin Environ Sustainability* 2:317–325.
 542 doi.org/10.1016/j.cosust.2010.09.002

543 Daan, H. (1985) Sensitivity of Verification Scores to the Classification of the Preditand. *Mon*
 544 *Weather Rev* 113:1384–1392

545 Daget, N., A.T. Weaver and M.A. Balmaseda, 2009: An ensemble three-dimensional
 546 variational data assimilation system for the global ocean: sensitivity to the
 547 observation- and background-error variance formulation. *Quart. J. Roy. Meteor. Soc.*,
 548 135, 1071-1094.

549 Dee, D.P. et al. (2011) The ERA-Interim reanalysis: Configuration and performance of the
 550 data assimilation system. *Q J R Meteorol Soc* 137:553–597

551 Déqué, M., et al. (1999), ARPEGE version 3, documentation algorithmique et mode d'emploi
 552 (in French), CNRM/GMGEC, Toulouse, France.

553 Doblas-Reyes, F.J. et al. (2013) Seasonal climate predictability and forecasting: Status and
 554 prospects. *Wiley Interdiscip Rev Clim Change* 4:245–268. doi:10.1002/wcc.217

555 Doblas-Reyes, F.J., Hagedorn, R. and Palmer, T.N. (2005) The rationale behind the success
 556 of multi model ensembles in seasonal forecasting–II. Calibration and combination.
 557 *Tellus A* 57:234–252

558 Ebinger, J. and Vergara, W. (2011) Climate Impacts on Energy Systems: Key Issues for
 559 Energy Sector Adaptation. World Bank.
 560 <https://openknowledge.worldbank.org/handle/10986/2271>
 561 Epstein, E.S. (1969) A Scoring System for Probability Forecasts of Ranked Categories. J
 562 Appl Meteorol 8:985–987
 563 Feddersen, H., A. Navarra, and M.N. Ward (1999) Reduction of model systematic error by
 564 statistical correction for dynamical seasonal predictions. J Clim 12:1974–1989
 565 Ferro, C. A. T., Richardson, D.S. and Weigel, A.P. (2008) On the effect of ensemble size on
 566 the discrete and continuous ranked probability scores. Meteorol Appl 15:19–24.
 567 Ferro, CAT (2014) Fair scores for ensemble forecasts. Quart. J. Roy. Meteor. Soc., 140,
 568 1917–1923.
 569 Foley, A.M. et al. (2012) Current methods and advances in forecasting of wind power
 570 generation. Renewable Energy, 37:1–8.
 571 <http://dx.doi.org/10.1016/j.renene.2011.05.033>.
 572 Frankfurt School-UNEP Collaborating Centre (2016) Global Trends in Renewable Energy
 573 Investment 2016. 1-84. Available at: [http://fs-unep-centre.org/publications/global-](http://fs-unep-centre.org/publications/global-trends-renewable-energy-investment-2016)
 574 [trends-renewable-energy-investment-2016](http://fs-unep-centre.org/publications/global-trends-renewable-energy-investment-2016)
 575 Hagedorn, R., Doblas-Reyes, F.J. & Palmer, T.N. (2005) The rationale behind the success of
 576 multi-model ensembles in seasonal forecasting - I. Basic concept. Tellus A 57:219–
 577 233
 578 IPCC (2012) Renewable Energy Sources and Climate Change Mitigation: Special Report of
 579 the Intergovernmental Panel on Climate Change, Cambridge University Press
 580 Jeong HI, Lee DY, Ashkok K, Ahn JB, Lee JY, Luo JJ, Schemm JK, Hendon HH, Braganza
 581 K, Ham YG (2012) Assessment of the APCC coupled MME suite in predicting the

582 distinctive climate impacts of two flavors of ENSO during boreal winter. *Clim Dyn*
 583 39:475-493.

584 Jeong HI, Ahn JB, Lee JY, Alessandri A, Hendon HH (2015) Interdecadal change of
 585 interannual variability and predictability of two types of ENSO. *Clim Dyn* 44: 1073-
 586 1091.

587 Johnson, C. and Bowler, N (2009) On the Reliability and Calibration of Ensemble Forecasts.
 588 *Mon Weather Rev* 137:1717–1720

589 Jolliffe, I.T. and Stephenson, D.B. (2003) Forecast verification: a practitioner's guide in
 590 atmospheric science. Wiley, ISBN: 0-471-49759-2

591 Kharin, V. V. and Zwiers, F.W. (2002) Climate predictions with multimodel ensembles. *J*
 592 *Clim* 15:793–799

593 Kharin, V. V., F. W. Zwiers, Q. Teng, G. J. Boer, J. Derome, and J. S. Fontecilla (2009) Skill
 594 assessment of seasonal hindcasts from the Canadian Historical Forecast Project.
 595 *Atmos Ocean* 47: 204–223

596 Kirtman BP, Min D, Infanti JM, et al (2014) The North American Multimodel Ensemble:
 597 Phase-1 Seasonal-to-Interannual Prediction; Phase-2 toward Developing
 598 Intraseasonal Prediction. *Bull Am Meteorol Soc* 95:585–601. doi: 10.1175/BAMS-
 599 D-12-00050.1

600 Koletsis, I., V. Kotroni, K.Lagouvardos, T.Soukissian (2016) Assessment of offshore wind
 601 speed and power potential over the Mediterranean and the Black Seas under future
 602 climate changes. *Renewable and Sustainable Energy Reviews* 60:234–245

603 Krishnamurti, T.N. et al. (2000) Multimodel Ensemble Forecasts for Weather and Seasonal
 604 Climate. *J Clim* 13:4196–4216

605 Kug, J.S., J.Y. Lee, and I.S. Kang (2008) Systematic error correction of dynamical seasonal
 606 prediction of sea surface temperature using a stepwise pattern project method. *Mon*
 607 *Weather Rev* 136:3501–3512. doi: 10.1175/2008MWR2272.1

608 Langford, S. and Hendon H.H. (2013) Improving Reliability of Coupled Model Forecasts of
 609 Australian Seasonal Rainfall. *Mon Weather Rev* 141:728–741

610 Lee, D. Y., J.-B. Ahn, and J.-H. Yoo (2015) Enhancement of seasonal prediction of East
 611 Asian summer rainfall related to western tropical Pacific convection. *Clim Dyn*
 612 45:1025-1042

613 Lee, D.Y., Ahn, J.B., et al. (2013) Improvement of grand multi-model ensemble prediction
 614 skills for the coupled models of APCC/ENSEMBLES using a climate filter. *Atmos*
 615 *Sci Lett* 14:139–145. doi:10.1002/asl2.430

616 Lee, D.Y., Ashok, K. and Ahn, J.B. (2011) Toward enhancement of prediction skills of
 617 multimodel ensemble seasonal prediction: a climate filter concept. *J Geophys Res*
 618 116:D06116. doi:10.1029/2010JD014610

619 Leung, L.R. et al. (1999) Simulations of the ENSO Hydroclimate Signals in the Pacific
 620 Northwest Columbia River Basin. *Bull Am Meteorol Soc* 80:2313–2329

621 Madec, G., P. Delecluse, M. Imbard, and C. Levy, 1998: Opa 8 ocean general circulation
 622 model -reference manual. Tech. rep., LODYC/IPSL Note 11

623 Meteo France (2015a) Météo-France seasonal forecast system 5 for Eurosip: Technical
 624 description. 1-38

625 Meteo France (2015b) Météo-France seasonal forecast system 5 versus system 4: Robust
 626 scores. 1-5

627 Michaelsen, J. (1987) Cross-Validation in Statistical Climate Forecast Models. *J Clim Appl*
 628 *Meteorol* 26:1589–1600

629 Min, Y.-M., Kryjov, V.N. and Park, C.-K. (2009) A Probabilistic Multimodel Ensemble
 630 Approach to Seasonal Prediction. *Weather Forecast* 24:812–828.
 631 <https://doi.org/10.1175/2008WAF2222140.1>

632 Molteni, F. et al. (2011) The new ECMWF seasonal forecast system (System 4), ECMWF
 633 Technical Memoranda, No. 656

634 Morse, A.P. et al. (2005) A forecast quality assessment of an end-to-end probabilistic multi-
 635 model seasonal forecast system using a malaria model. *Tellus A* 57:464–475

636 Murphy, A.H. (1971) A Note on the Ranked Probability Score. *J Appl Meteorol* 10:155–156

637 Murphy, A.H. (1988) Skill scores based on the Mean square error and their relationships to
 638 the correlation coefficient. *Mon Weather Rev* 116:2417–2424

639 Palmer, B.T.N. (2000) A probability and decision-model analysis of PROVOST seasonal
 640 multi-model ensemble integrations. *Q J R Meteorol Soc* 126:2013–2033

641 Palmer, T.N. et al. (2005) Probabilistic prediction of climate using multi-model ensembles:
 642 from basics to applications. *Philos Trans R Soc Lond B Biol Sci* 360:1991–1998

643 Pan, J. and H. Van den Dool (1998) Extended-range probability forecasts based on dynamical
 644 model output. *Weather Forecast* 13:983–996

645 Pavan, V. and Doblas-Reyes, F.J. (2000) Multi-model seasonal hindcasts over the Euro-
 646 Atlantic: skill scores and dynamic features. *Clim Dyn* 16:611–625.
 647 <http://doi.org/10.1007/s003820000063>

648 Peng, P. et al. (2002) An analysis of multimodel ensemble predictions for seasonal climate
 649 anomalies. *J Geophys Res* 107:1–12. doi:10.1029/2002JD002712

650 Richardson, D.S. (2001) Measures of skill and value of ensemble prediction systems, their
 651 interrelationship and the effect of ensemble size. *Q J R Meteorol Soc* 127:2473–
 652 2489

653 Royer JF, Cariolle D, Chauvin F, De'que' M, Douville H, Hu RM, Planton S, Rascol A,
 654 Ricard JL, Salas y Me'lia D, Sevault F, Simon P, Somot S, Tyteca S, Terray L, Valcke
 655 S (2002) Simulation des changements climatiques au cours du 21-e' sie'cle
 656 incluant l'ozone stratosphe'rique (simulation of climate changes during the 21-st
 657 century including stratospheric ozone). *C R Geosci* 334:147–154
 658 Torralba, V. et al. (2017) Seasonal climate prediction: a new source of information for the
 659 management of wind energy resources. *J Appl Meteorol Clim* 56:1231–1247
 660 <http://doi.org/10.1175/JAMC-D-16-0204.1>
 661 Troccoli, A. et al. (2013) Promoting new links between energy and meteorology. *Bull Am*
 662 *Meteorol Soc* 94: ES36–ES40. <https://doi.org/10.1175/BAMS-D-12-00061.1>
 663 Troccoli, A. et al. (2010) Weather and climate risk management in the energy sector. *Bull Am*
 664 *Meteorol Soc* 91:785–788
 665 Vladislavleva, E. et al. (2013) Predicting the energy output of wind farms based on weather
 666 data: Important variables and their correlation. *Renewable Energy* 50:236–243
 667 Voldoire, A. et al. (2013) The CNRM-CM5.1 global climate model: Description and basic
 668 evaluation. *Clim Dyn* 40:2091–2121
 669 Wang, B. et al. (2008) How accurately do coupled climate models predict the leading modes
 670 of Asian-Australian monsoon interannual variability? *Clim Dyn* 30:605–619
 671 Weigel, A.P. et al. (2010) Risks of model weighting in multimodel climate projections. *J Clim*
 672 23:4175–4191
 673 Weigel, A.P., Liniger, M.A. and Appenzeller, C. (2008) Can multi-model combination really
 674 enhance the prediction skill of probabilistic ensemble forecasts? *Q J R Meteorol Soc*
 675 134:241–260

- Weigel, A.P., Liniger, M.A. & Appenzeller, C. (2007) The Discrete Brier and Ranked Probability Skill Scores. *Mon Weather Rev* 135:118–124
<https://doi.org/10.1175/MWR3280.1>.
- Weisheimer, A., and Coauthors, 2009: ENSEMBLES: A new multimodel ensemble for seasonal-to-annual predictions—Skill and progress beyond DEMETER in forecasting tropical Pacific SSTs. *Geophys. Res. Lett.*, 36, L21711, doi:10.1029/2009GL040896.
- Wilks, D.S. (2006) Statistical methods in the atmospheric sciences. Academic Press, 627pp, ISSN 0074-6142
- Yang, D., X.-Q. Yang, Q. Xie, Y. Zhang, X. Ren, and Y. Tang (2016) Probabilistic versus deterministic skill in predicting the western North Pacific-East Asian summer monsoon variability with multimodel ensembles. *J Geophys Res* 121:1079–1103. doi:10.1002/ 2015JD023781
- Yun, W.T. et al. (2005) A multi-model superensemble algorithm for seasonal climate prediction using DEMETER forecasts. *Tellus A* 57:280–289

List of Table Title

Table 1 Description of the coupled global seasonal prediction systems employed

List of Figure Captions

Fig. 1 Temporal correlation coefficients (TCCs) between the ERA-Interim and ensemble mean forecasts (a-b: MMEs and c-f: individual models) for 10m wind speed during boreal winter (December through February, DJF) for period 1991-2012. (a) MME4 and (b) MME2 are the multi-model ensemble predictions using the total of four models and the two (c and f) better performing models, respectively. Hatched areas highlight regions where TCC is significant at the 90% confidence level from a two-tailed Student's t-test. The upper right values of each map are the area-averaged TCCs.

Fig. 2 Fair ranked probability skill score (FRPSS) for tercile events of 10-m wind speed from (a-b) MMEs and (c-f) individual models with respect to the ERA-Interim reference climatology during winter (DJF) for period 1991-2012. Hatched areas highlight regions where FRPSS is significant at the 95% confidence level from a one-tailed Z-test.

Fig. 3 Temporal correlation coefficients (TCCs) between the ERA-Interim and ensemble mean forecasts (left and right columns: bias-adjusted MME4 and MME2) for 10m wind speed during boreal winter (December through February, DJF) for period 1991-2012. Upper and lower rows show the skill scores for (a-b) simple bias corrected (SBC) and (c-d) calibrated (Cal) MME predictions, respectively. Hatched areas highlight regions where TCC is significant at the 90% confidence level from a two-tailed Student's t-test.

Fig. 4 Fair ranked probability skill score (FRPSS) for tercile events of 10-m wind speed from bias-adjusted MME4 (left column) and MME2 (right column) predictions with respect to the ERA-Interim reference climatology during winter (DJF) for period 1991-2012. Upper and lower rows show the skill scores for (a-b) simple bias corrected (SBC) and (c-d) calibrated (Cal) MME predictions, respectively. Hatched areas highlight regions where FRPSS is significant at the 95% confidence level from a one-tailed Z-test.

Fig. 5 Reliability diagrams (lines) for probabilistic categorical forecasts (tercile events) of global 10m wind speed in terms of MME4 (left column) and MME2 (right column) predictions obtained by the simple bias correction (SBC, red) and calibration (Cal, blue) method. Upper and lower rows correspond to above and below normal categories, respectively. Vertical color bars on the diagonal within the reliability diagrams depict consistency bars for a 95% confidence level in each bin. The sharpness diagrams (bars) at the right of the reliability diagrams represent the relative frequency distributions of the probability forecasts.

Fig. 6 Same as Fig. 5, except for Northern Europe (15°W-45°E, 45°N-75°N) region.

Fig. 7 Fair ranked probability skill score (FRPSS) for tercile events of 10-m wind speed from the MME2 raw predictions with respect to the ERA-Interim reference climatology during four seasons for period 1991-2012. Hatched areas highlight regions where FRPSS is significant at the 95% confidence level from a one-tailed Z-test.

Fig. S1 Root mean square skill score (RMSSS) of the MME4 (left column) and MME2 (right column) predictions with respect to the ERA-Interim reference climatology for 10m wind speed during winter (DJF) for period 1991-2012. Upper, middle and lower rows show the skill scores for (a-b) raw, (c-d) simple bias corrected (SBC) and (e-f)

calibrated (Cal) MME predictions, respectively. Hatched areas highlight regions where RMSSS is significant at the 95% confidence level from a one-tailed F-test.

Fig. S2 Reliability diagrams (lines) for probabilistic categorical forecasts (tercile events) of global 10m wind speed in terms of raw predictions of individual models and MME4. (a) Left and (b) Right panels correspond to above and below normal categories, respectively. Vertical color bars on the diagonal within the reliability diagrams depict consistency bars for a 95% confidence level in each bin. The sharpness diagrams (bars) at the right of the reliability diagrams represent the relative frequency distributions of the probability forecasts.

Fig. S3 Temporal correlation coefficients (TCCs) between the ERA-Interim and ensemble mean forecasts from the MME2 raw predictions for 10m wind speed during four seasons for period 1991-2012. Hatched areas highlight regions where TCC is significant at the 90% confidence level from a two-tailed Student's t-test.

754

755 Table 1

Model Name	Atmospheric Model	Resolution	Oceanic Model	Resolution	Ensemble size of the hindcasts
ECMWF-S4	IFS CY36R4	TL255L91	NEMO3.0	1°lat x 1°lon L42	51
METFR-S3	ARPEGE4.6	T63L31	OPA8.2	2°lat x 2°lon L31	11
METFR-S4	ARPEGE5.2	TL127L31	NEMO3.2	1°lat x 1°lon L42	15
METFR-S5	ARPEGE6.1	T255L91	NEMO3.2	1°lat x 1°lon L42	15

756

757 **Figures**

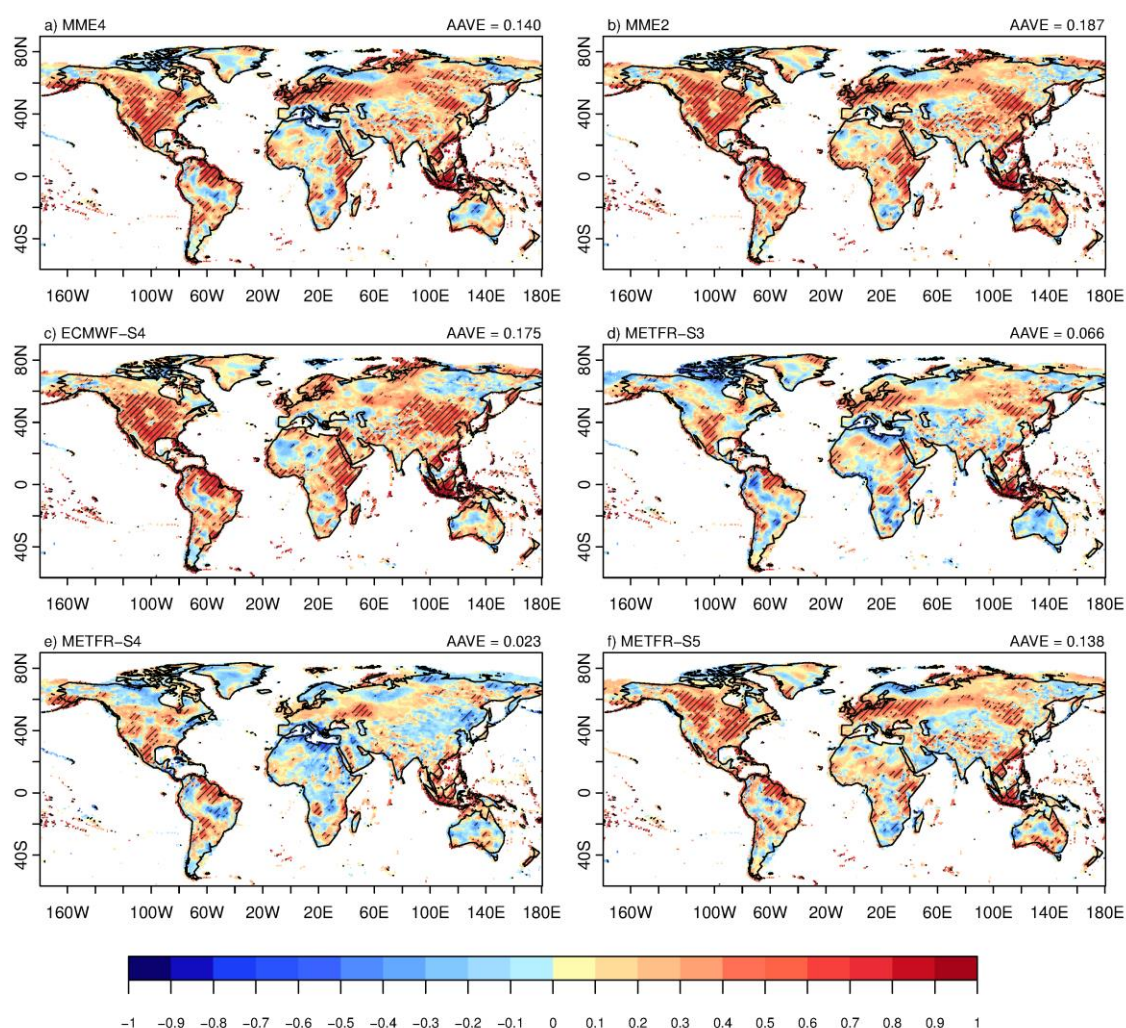


Fig. 1 Temporal correlation coefficients (TCCs) between the ERA-Interim and ensemble mean forecasts (a-b: MMEs and c-f: individual models) for 10m wind speed during boreal winter (December through February, DJF) for period 1991-2012. (a) MME4 and (b) MME2 are the multi-model ensemble predictions using the total of four models and the two (c and f) better performing models, respectively. Hatched areas highlight regions where TCC is significant at the 90% confidence level from a two-tailed Student's t-test. The upper right values of each map are the area-averaged TCCs.

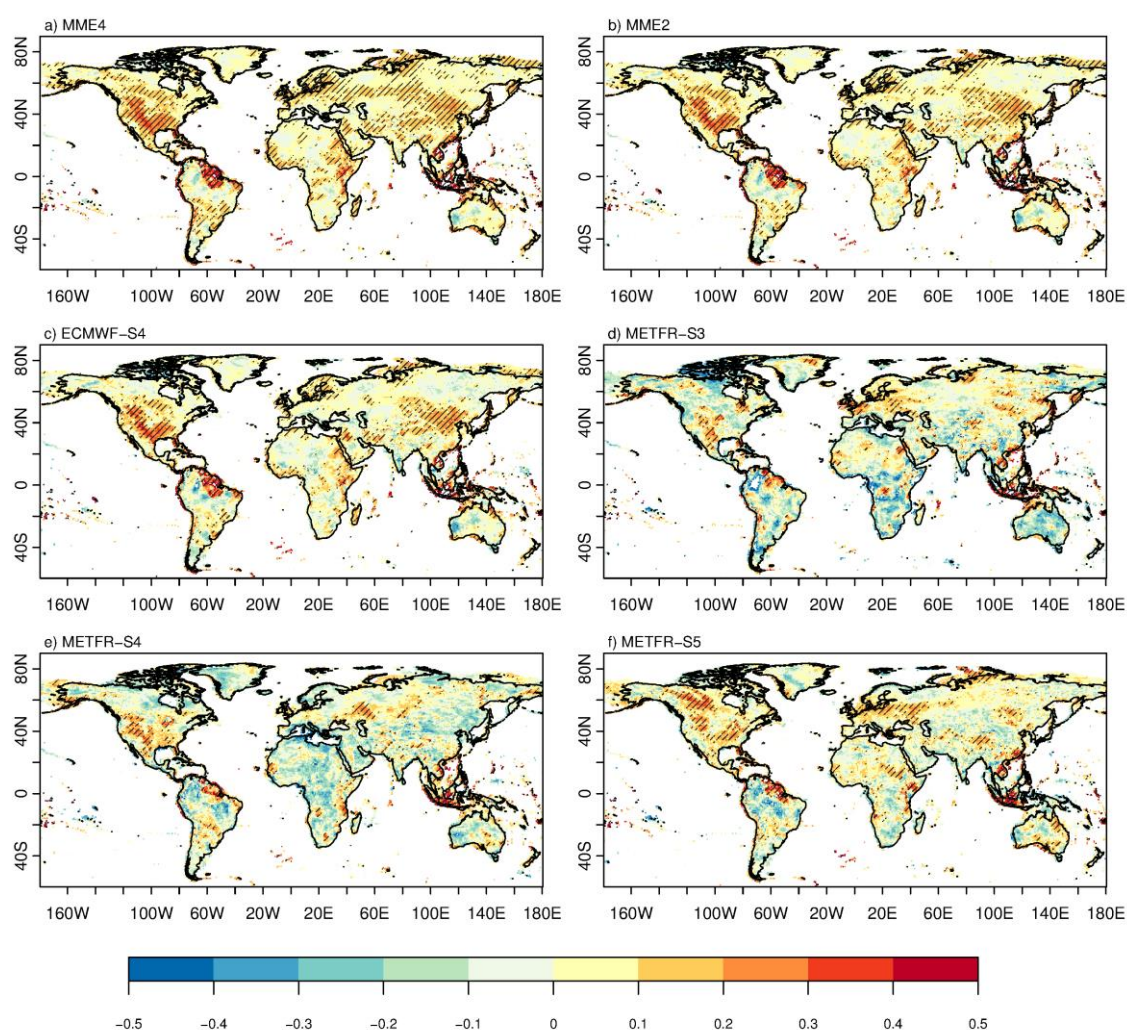


Fig. 2 Fair ranked probability skill score (FRPSS) for tercile events of 10-m wind speed from (a-b) MMEs and (c-f) individual models with respect to the ERA-Interim reference climatology during winter (DJF) for period 1991-2012. Hatched areas highlight regions where FRPSS is significant at the 95% confidence level from a one-tailed Z-test.

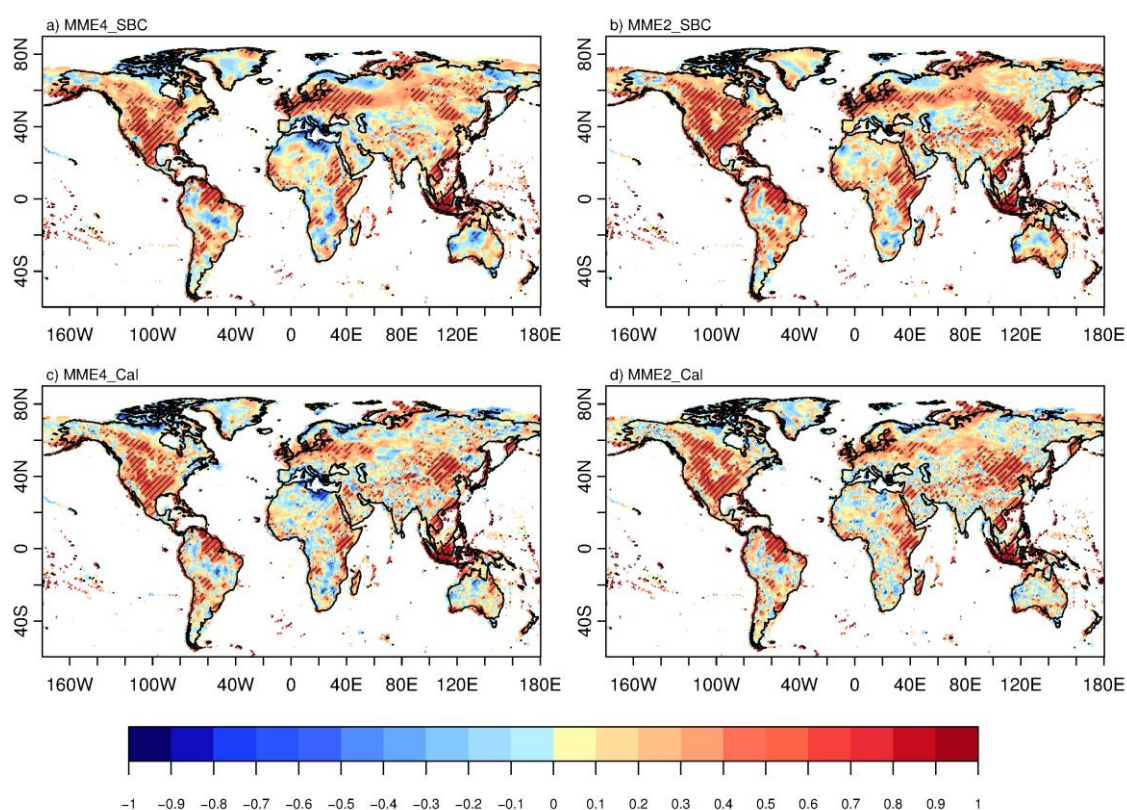


Fig. 3 Temporal correlation coefficients (TCCs) between the ERA-Interim and ensemble mean forecasts (left and right columns: bias-adjusted MME4 and MME2) for 10m wind speed during boreal winter (December through February, DJF) for period 1991-2012. Upper and lower rows show the skill scores for (a-b) simple bias corrected (SBC) and (c-d) calibrated (Cal) MME predictions, respectively. Hatched areas highlight regions where TCC is significant at the 90% confidence level from a two-tailed Student's t-test.

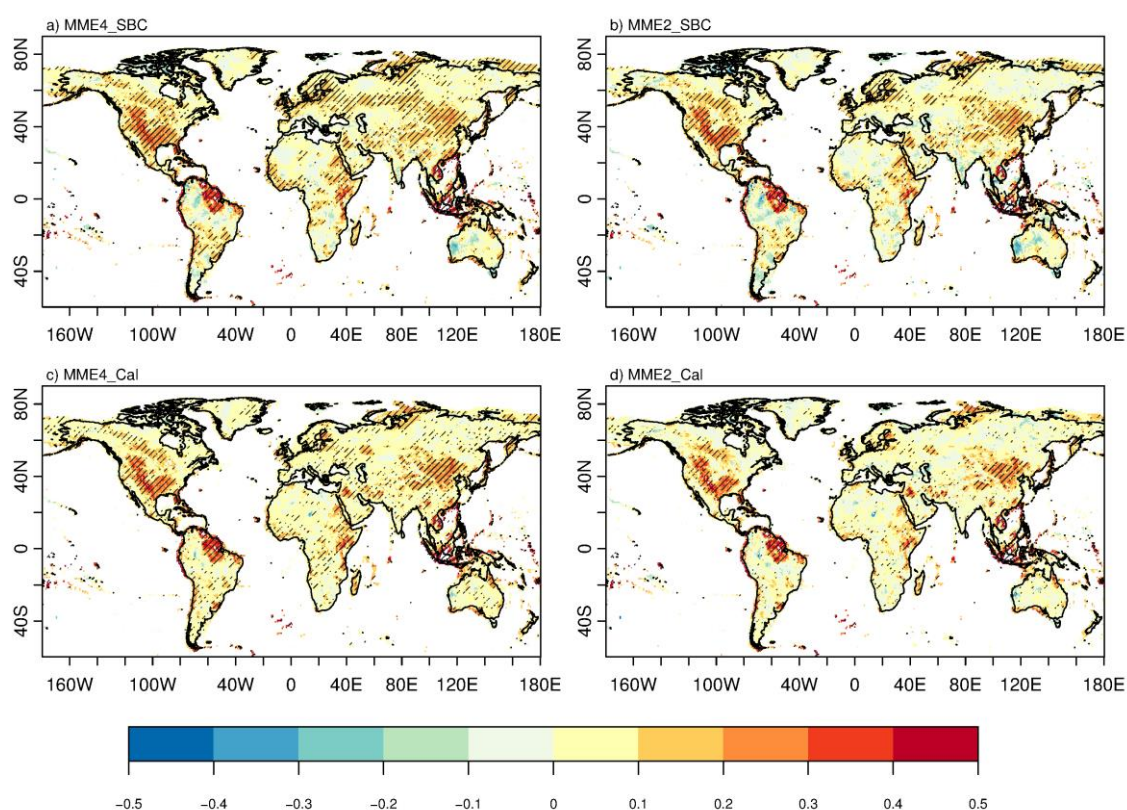


Fig. 4 Fair ranked probability skill score (FRPSS) for tercile events of 10-m wind speed from bias-adjusted MME4 (left column) and MME2 (right column) predictions with respect to the ERA-Interim reference climatology during winter (DJF) for period 1991-2012. Upper and lower rows show the skill scores for (a-b) simple bias corrected (SBC) and (c-d) calibrated (Cal) MME predictions, respectively. Hatched areas highlight regions where FRPSS is significant at the 95% confidence level from a one-tailed Z-test.

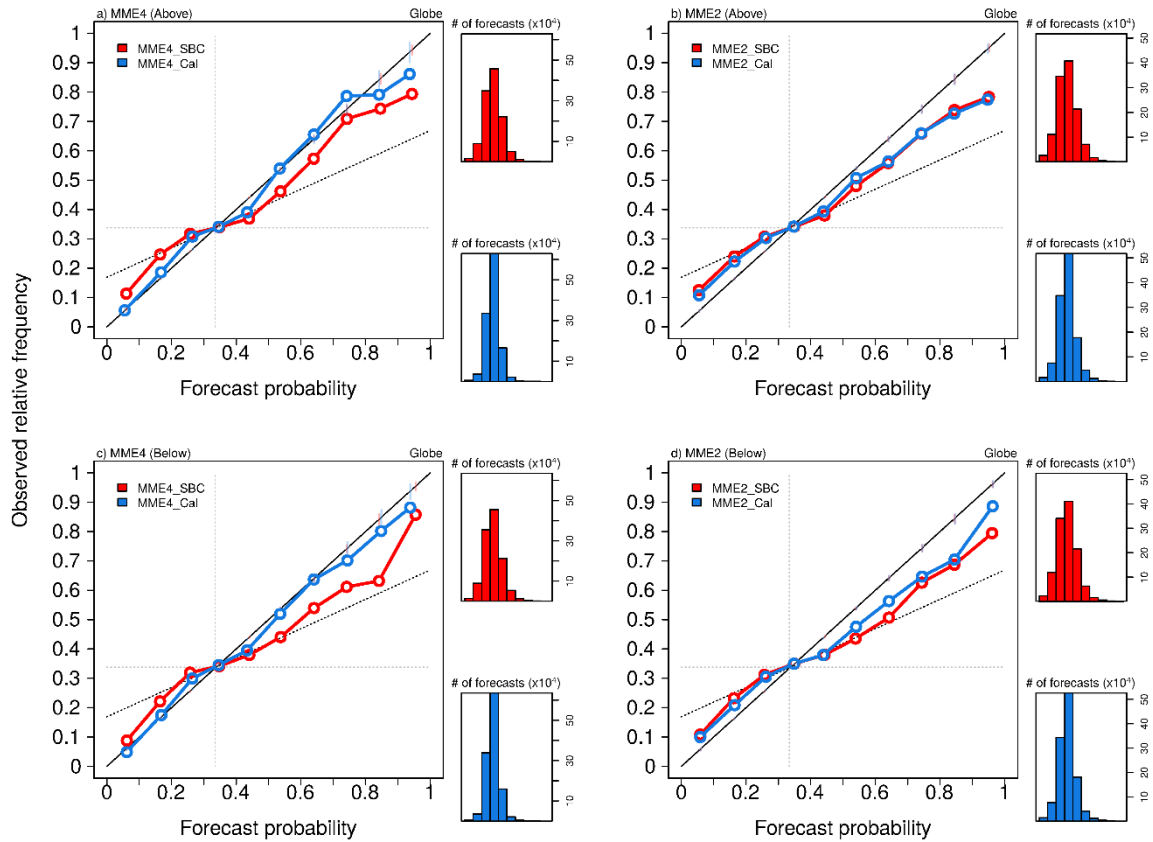


Fig. 5 Reliability diagrams (lines) for probabilistic categorical forecasts (tercile events) of global 10m wind speed in terms of MME4 (left column) and MME2 (right column) predictions obtained by the simple bias correction (SBC, red) and calibration (Cal, blue) method. Upper and lower rows correspond to above and below normal categories, respectively. Vertical color bars on the diagonal within the reliability diagrams depict consistency bars for a 95% confidence level in each bin. The sharpness diagrams (bars) at the right of the reliability diagrams represent the relative frequency distributions of the probability forecasts.

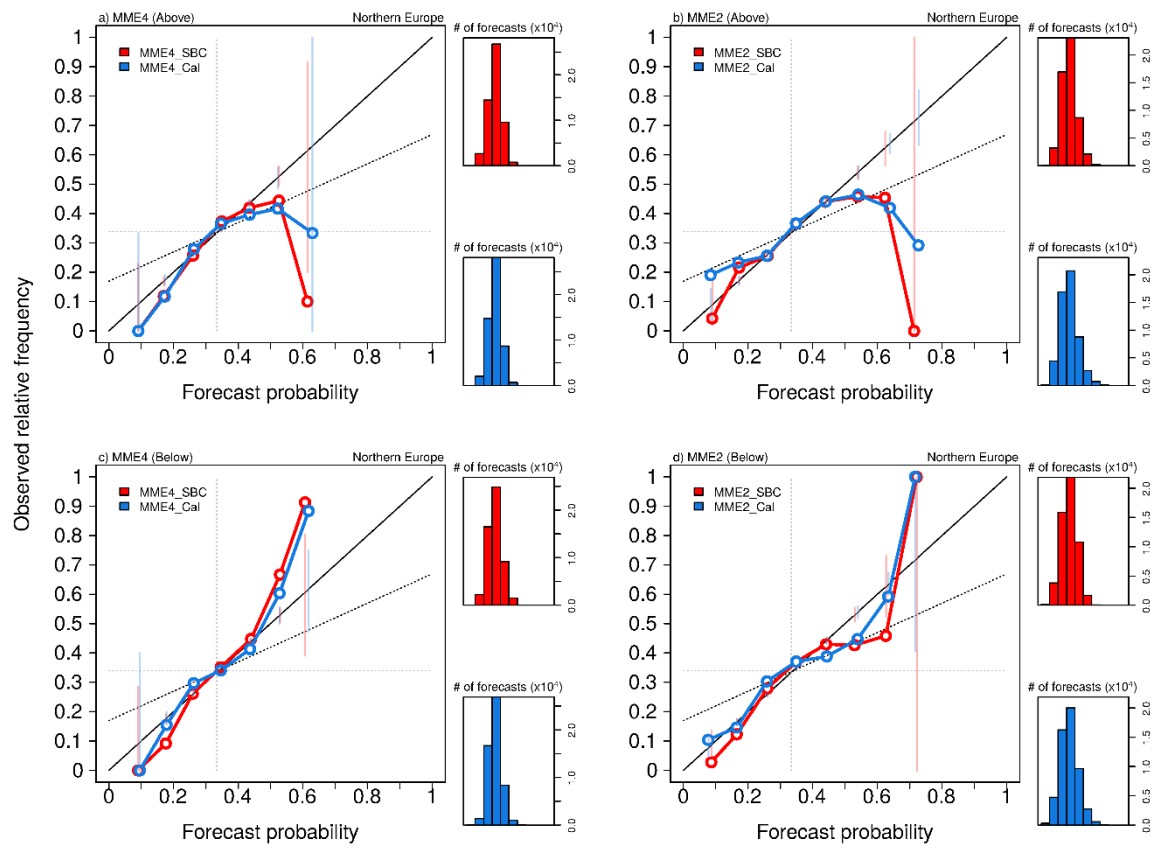


Fig. 6 Same as Fig. 5, except for Northern Europe (15°W-45°E, 45°N-75°N) region.

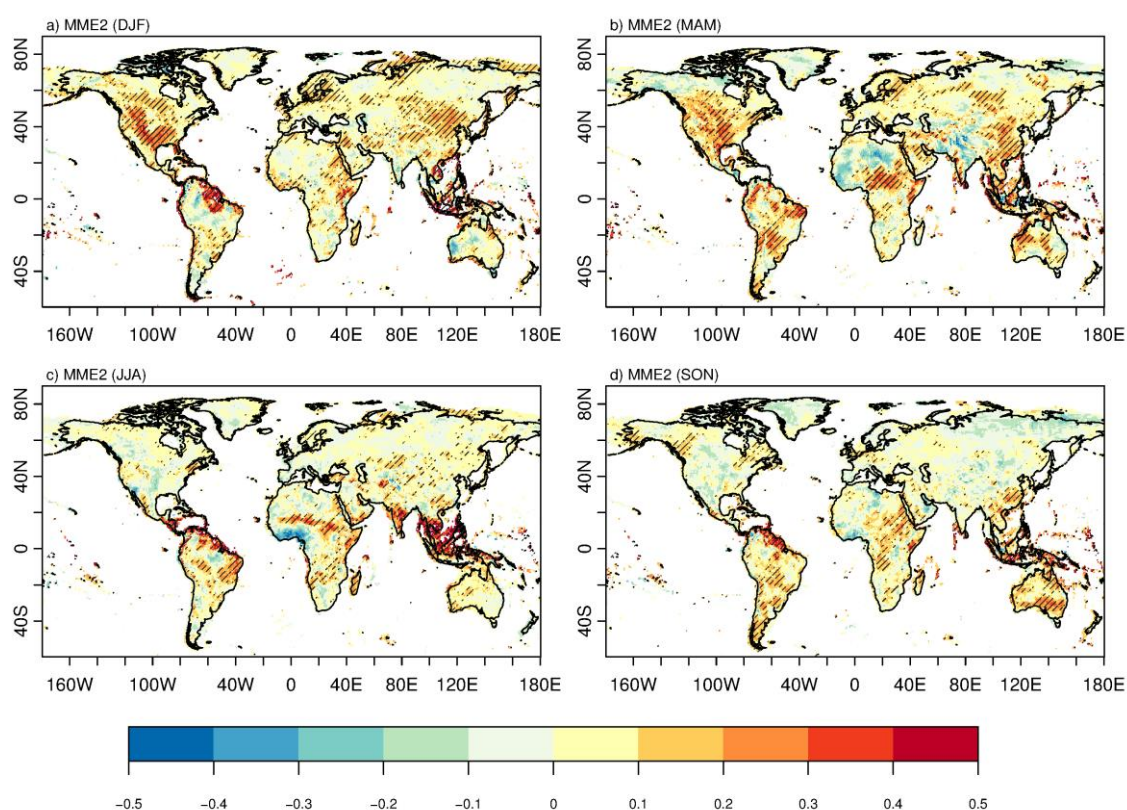


Fig. 7 Fair ranked probability skill score (FRPSS) for tercile events of 10-m wind speed from the MME2 raw predictions with respect to the ERA-Interim reference climatology during four seasons for period 1991-2012. Hatched areas highlight regions where FRPSS is significant at the 95% confidence level from a one-tailed Z-test.

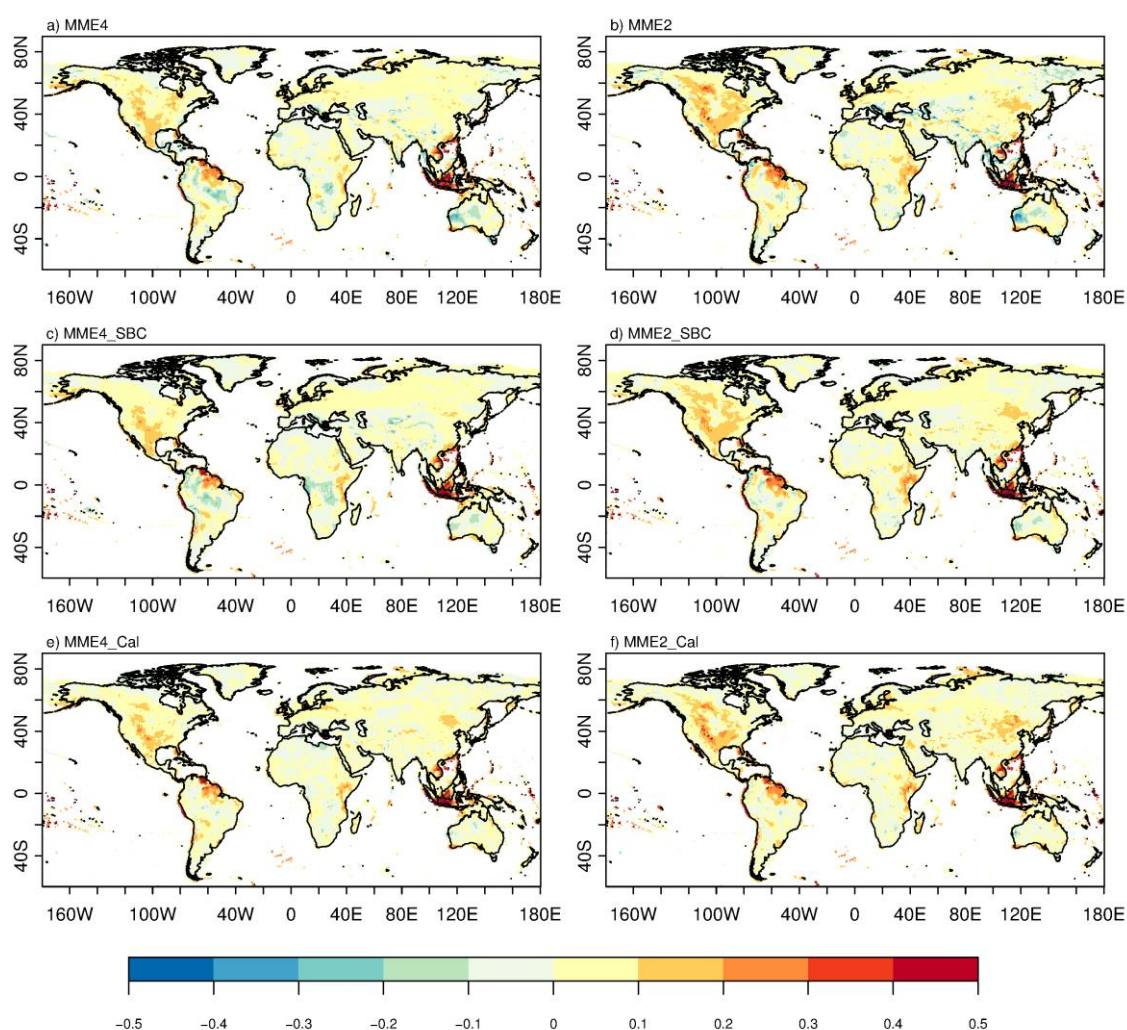


Fig. S1 Root mean square skill score (RMSSS) of the MME4 (left column) and MME2 (right column) predictions with respect to the ERA-Interim reference climatology for 10m wind speed during winter (DJF) for period 1991-2012. Upper, middle and lower rows show the skill scores for (a-b) raw, (c-d) simple bias corrected (SBC) and (e-f) calibrated (Cal) MME predictions, respectively. Hatched areas highlight regions where RMSSS is significant at the 95% confidence level from a one-tailed F-test.

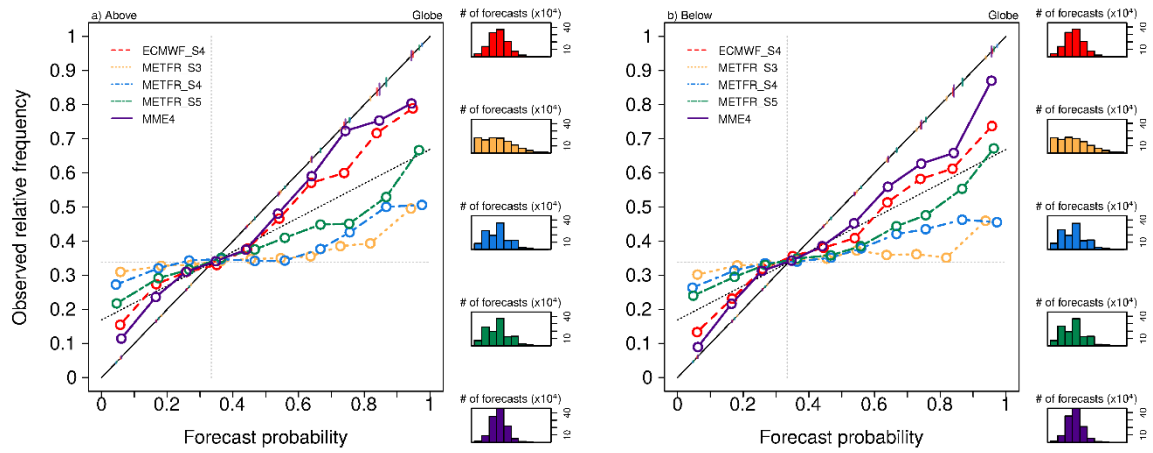


Fig. S2 Reliability diagrams (lines) for probabilistic categorical forecasts (tercile events) of global 10m wind speed in terms of raw predictions of individual models and MME4. (a) Left and (b) Right panels correspond to above and below normal categories, respectively. Vertical color bars on the diagonal within the reliability diagrams depict consistency bars for a 95% confidence level in each bin. The sharpness diagrams (bars) at the right of the reliability diagrams represent the relative frequency distributions of the probability forecasts.

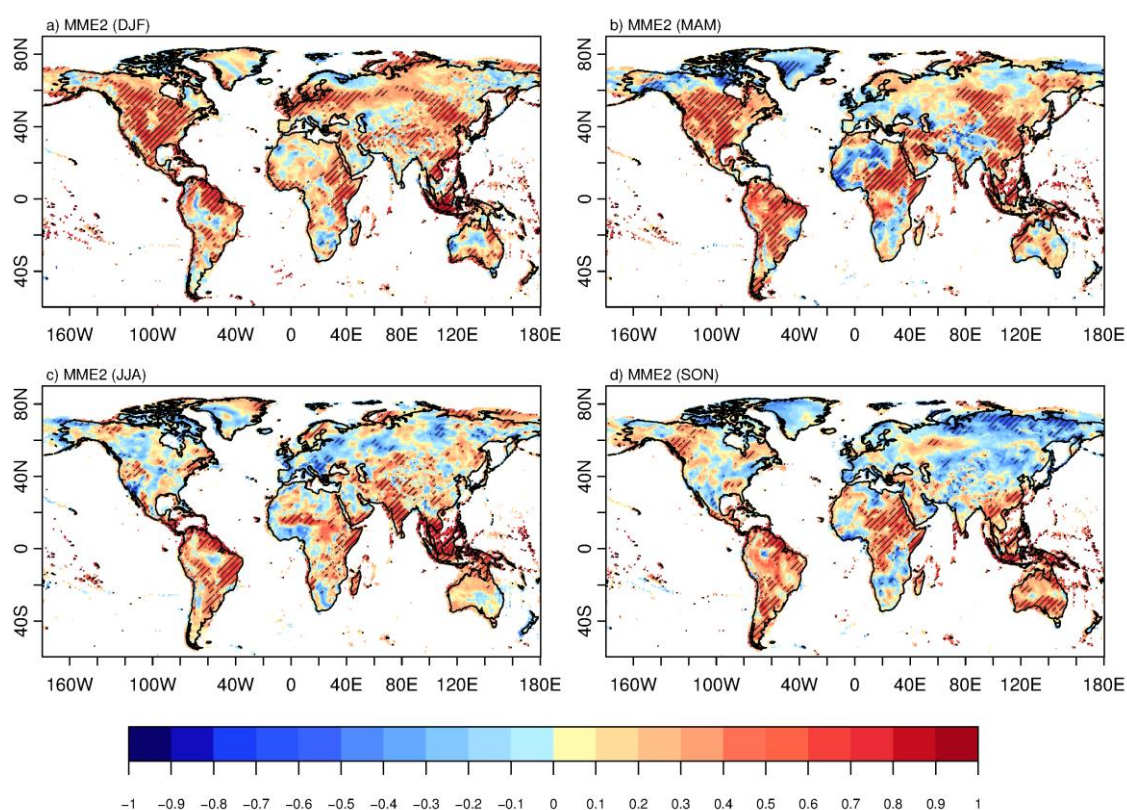


Fig. S3 Temporal correlation coefficients (TCCs) between the ERA-Interim and ensemble mean forecasts from the MME2 raw predictions for 10m wind speed during four seasons for period 1991-2012. Hatched areas highlight regions where TCC is significant at the 90% confidence level from a two-tailed Student's t-test.